国　　际　　电　　信　　联　　盟

ITU-R

国际电联无线电通信部门

ITU-R BS. 1534-2 建议书

(06/2014)

音频系统中级质量水平的主观评价方法

BS 系列

广播业务（声音）

1865–2015

国际电信联盟

前言

无线电通信部门的职责是确保卫星业务等所有无线电通信业务合理、平等、有效、经济地使用无线电频谱，不受频率范围限制地开展研究并在此基础上通过建议书。

无线电通信部门的规则和政策职能由世界或区域无线电通信大会以及无线电通信全会在研究组的支持下履行。

知识产权政策（IPR）

ITU-R的IPR政策述于ITU-R第1号决议的附件1中所参引的《ITU-T/ITU-R/ISO/IEC的通用专利政策》。专利持有人用于提交专利声明和许可声明的表格可从http://www.itu.int/ITU-R/go/patents/en获得，在此处也可获取《ITU-T/ITU-R/ISO/IEC的通用专利政策实施指南》和ITU-R专利信息数据库。

ITU-R 系列建议书

（也可在线查询 http://www.itu.int/publ/R-REC/en）

系列	标题
BO	卫星传送
BR	用于制作、存档和播出的录制；电视电影
BS	**广播业务（声音）**
BT	广播业务（电视）
F	固定业务
M	移动、无线电定位、业余和相关卫星业务
P	无线电波传播
RA	射电天文
RS	遥感系统
S	卫星固定业务
SA	空间应用和气象
SF	卫星固定业务和固定业务系统间的频率共用和协调
SM	频谱管理
SNG	卫星新闻采集
TF	时间信号和频率标准发射
V	词汇和相关问题

说明：该ITU-R建议书的英文版本根据ITU-R第1号决议详述的程序予以批准。

电子出版
2015年，日内瓦

ITU-R BS.1534-2 建议书

音频系统中级质量水平的主观评价方法

(ITU-R第62/6号研究课题)

(2001-2003-2014年)

范围

本建议书描述了一种用于对中级音频质量进行主观评价的方法。此方法映射了很多ITU-R BS.1116的概念，并且使用了与对图片质量进行评价(即，ITU-R BT.500建议书)所用相同的等级量表。

已经成功地测试了被称为“带有隐藏参考和锚点的多激励测试 (MUSHRA)”的方法。这些测试显示出MUSHRA方法适合于对中级音频质量进行评价，并给出准确和可靠的结果。

关键词

聆听测试、伪像、中级音频质量、音频编码、主观评价、音频质量。

国际电联无线电通信全会，

考虑到

a) ITU-R BS.1116、ITU-R BS.1284、ITU-R BT.500、ITU-R BT.710和ITU-R BT.811建议书以及ITU-T P.800、ITU-T P.810和ITU-T P.830建议书已经确立了评价音频、视频和语音系统主观质量的方法，

b) 新型传送服务可以中级音频质量运行，例如：互联网或固态播放器上的流音频、数字卫星业务、数字短波和中波系统或移动多媒体应用；

c) ITU-R BS.1116建议书的目的是用于对小损伤进行评价，不适合于对中级音频质量系统的评价；

d) ITU-R BS.1284建议书不提供对中级音频质量评价的绝对评分；

e) 测试中包括适当和相关的锚点可以实现主观评分量表的稳定应用；

f) ITU-T P.800、ITU-T P.810和ITU-T P.830建议书专注于在电话环境中的语音信号，并且已证明不足以用于对广播环境中的音频信号进行评价；

g) 使用标准化的主观测试方法对测试数据的交换、兼容性和纠正评价是重要的；

h) 新的多媒体业务可能需要对音频和视频质量的组合进行评价；

i) 名称MUSHRA经常被误用于不使用参考和锚点的测试；

j) 锚点可以影响测试的结果，而且令人满意的是锚点类似于被测系统的伪像，

建议

1 应该将本建议书附录1中给出的测试和评价程序用于对中级音频质量的主观评价，

进一步建议

1 继续对在最先进音频系统中所遇到的具有损伤特性的锚点进行研究，并且应该更新本建议书来包括适当的新锚点。

附件 1

1 引言

本建议书描述了一种对中级音频质量进行主观评价的方法。此方法映射了很多ITU-R BS.1116的概念，并且使用了与对图片质量进行评价所用相同的等级量表(即，ITU-R BT.500建议书)。

已经成功地测试了被称为“带有隐式参考和锚点的多激励测试 (MUSHRA)”的方法。这些测试显示出MUSHRA方法适合于对中级音频质量进行评价，并给出准确和可靠的结果，[2; 4; 3]。

本建议书包括以下章节和附件：

第1节： 引言

第2节： 新方法的范围、测试动机和目的

第3节： 实验设计

第4节： 选择评价者

第5节： 测试方法

第6节： 属性

第7节： 测试素材

第8节： 聆听条件

第9节： 统计分析

第10节： 测试报告和结果呈现

附件 1 (规范性) 对评价者的指导

附件 2 (参考性) 对用户界面设计的指导说明

附件 3 (规范性) 对采用再抽样技术和Monte-Carlo仿真方法在二个样本之间进行非参量统计比较的描述

附件 4 (参考性) 对参量统计分析的指导说明

附件 5 (参考性) 对最佳锚点特性的要求

2 新方法的范围、测试动机和目的

主观聆听测试被认为仍然是最可靠的音频系统质量测量方法。存在有经过很好描述和证明的对在质量范围顶部和底部音频质量进行评价的方法。

ITU-R BS.1116建议书 – 用于在包括多通道音响系统的音频系统中主观评价小损伤的方法，被用于对具有小损伤的高质量音频系统进行评价。但是，有的应用中较低质量音频是可接受或不可避免的。在互联网用于音频资料分送和广播中的快速发展已经导致了在数据速率受限情况下对音频质量的危害。可能包含中级音频质量的其他应用有数字AM (即，世界数字无线电 (DRM)、数字卫星广播、收音机和电视中的解说电路、音频点播服务和拨号线路上的音频)。在ITU-R BS.1116建议书中定义的测试方法不完全适合于评价这些较低质量的音频系统[4]，因为它对量表底部质量微小差别之间的区分较差。

ITU-R BS.1284建议书仅仅给出了专用于高质量音频范围的方法，或者未给出音频质量的绝对评分。

其他建议书，如ITU-T P.800、ITU-T P.810或ITU-T P.830，专注于对一个电话环境下语音信号的主观评价。欧洲广播联盟(EBU)项目组B/AIM已经使用这些ITU-T方法采用和一个广播环境下相同的典型音频素材进行了实验。这些方法中没有一个同时满足对一个绝对量表、与一个参考信号比较和在具有一个合理评价者数量的小置信区间的要求。因此，不能采用任何这些方法适当地完成在一个广播环境中对音频信号进行评价。

本建议书中所描述的改进测试方法的目的是提供一个可靠和可重复的系统测量，这些系统的音频质量通常会落入ITU-R BS.1116建议书所使用的损伤量表下半部分[2; 4; 3]。在MUSHRA测试方法中，使用一个高质量参考信号，而且认为被测系统将引入显著的损伤。MUSHRA将要用于对中级质量音频系统的评价。如果采用适当的内容来使用MUSHRA，理想的情况是收听者评分范围将在20-80 MUSHRA分之间。如果测试条件大部分的评分落入80-100的范围中，很可能测试结果是无效的。

压缩评分可能的原因是：使用了缺乏经验的评价者、使用了不利于评价的内容，或者对测试编码算法做了不适当的测试选择。

3 实验设计

很多不同种类的研究策略被用于在科学关注范围内收集可靠信息。在对音频系统中的损伤进行主观评价中，应该使用最正式的实验方法。主观实验的特征首先是对实验条件进行实际控制和操纵，其次是对来自收听者的统计数据进行收集和分析。需要精心的实验设计和规划来保证使可能在测试结果中造成不明确性的非受控因素最小。例如，如果在一个聆听测试中音频项实际序列对所有评价者相同，则无法确认评价者所做出的判定是由于这个序列而不是由于呈现出的不同损伤水平所引起。相应地，测试条件必须以揭示独立因素影响的方式来安排，并且仅仅是这些因素。

在可以预期潜在损伤和其他特性将均匀地分布在整个聆听测试中的情形下，可以对测试条件的呈现应用一个真正的随机化。在预期不均匀性的情况下，在呈现测试条件中必须对此

予以考虑。例如，当要评价的素材在难度水平上变化时，激励呈现的顺序在进程之内和之间都必须是随机分布。

需要设计聆听测试不使评价者超负荷到判断准确性降低点。除了声音和视觉之间关系很重要的情况之外，优选在没有伴随画面的情况下对音频系统进行评价。一个主要的考虑是包括适当的控制条件。通常，控制条件包括呈现无损伤音频素材，以评价者无法预测的方式引入。正是这些控制激励与潜在无损伤音频资料评价之间的差别使得可以得出这样的结论，分级就是对损伤的实际评价。

以下将描述一些这样的考虑。应该明白，实验设计的主题、实验实施和统计分析是复杂的，并且不是所有细节都能像这个建议书一样提供。建议在聆听测试规划开始时就应该咨询或者引入具有实验设计和统计专门知识的专业人员。

为了能够有效地在实验室之间分析和传递数据，应该报告实验设计。应该详细地定义相关的和独立的变量。独立变量的数量将用它们相关联的级别来确定。

4 选择评价者

如同在ITU-R BS.1116建议书中，来自评价音频系统小损伤的聆听测试的数据应该来自具有发现这些小损伤经验的评价者。这些要被测试的系统所达到的质量越高，具有经验丰富的收听者就越重要。

4.1 选择评价者的标准

尽管MUSHRA测试方法的目的不是要评价小的损伤，但是仍然建议应该采用有经验的收听者，以保证所采集测试数据的良好性。这些收听者应该具有以一种评价性的方式收听声音的经验。这样的收听者将比无经验的收听者更块地给出一个可靠的结果。同样重要的是要注意，大多数无经验的收听者在频繁面对之后趋于变得对各种类型的伪像更加敏感。一个有经验的评价者因其完成一个聆听测试的能力而被选定。根据对评估的重复，将按照评价者在一个测试中的可靠性和辨识技巧来决定这个能力是否合格并量被化，如下所定义：

– **辨识**：对感觉到测试项之间差别能力的一个度量。

– **可靠性**：对重复对相同测试项进行评分的接近程度的一个度量。

应该仅仅将对任何给定测试划分为*有经验的评价者*的那些评价者计入到最终的数据分析中。很多执行对评价者进行这种分析的技术是可用的。更多信息请参考ITU-R BS.2300报告[1]。这些是基于至少一个由每个评价者进行的重复评分，并且允许在一个实验内对评价者经验进行合格确认和量化。这些方法应该作为在一个先导实验内对评价者的预筛选，或者更可

[1] ITU-R BS.2300-0报告中所描述的专业知识度量(eGauge)法是这种技术实施的一个实例。它可以在http://www.itu.int/oth/R0A07000036得到。

取的是，既作为预筛选又作为主测试的一部分来应用。一个先导实验与一系列实验相关，并且包含要在主实验中评估的一个代表性测试样本集。出于对收听者专业知识进行评价的目的，先导实验应该包含测试激励的一个相关子集，代表在实际主实验期间要被评估的整个激励和伪像范围。

分析的图形显示应该表达出关于评价者可靠性相对于辨识力的信息。

4.1.1 对评价者的预筛选

聆听评价组应该由有经验的收听者组成，换句话说，是由懂得并对所描述的主观质量评估方法进行过适当培训的人组成。这些收听者应该：

– 具有以一种评价性的方式收听声音的经验；

– 具有正常的听力（应该将ISO标准389用作指导）。

培训程序应该被用作一种预筛选的工具。仅仅将在一个先导实验或主实验中被划分为有经验评价者的收听者计入数据分析中。

包含激励重复被用来提供一种评价收听者可靠性的方法。

对引入一种预筛选技术的主要理由是要增加聆听测试的效率。但是，这必须要针对过度限制结果关联性的风险进行平衡。

4.1.2 对评价者的后筛选

后筛选方法排除对一个明显损伤的锚点信号给予非常高评分的那些评价者和经常对隐藏参考评定为明显受损伤的那些评价者，如以下标准所定义：

– 如果一个评价者对> 15%测试项的隐藏参考条件评分低于90分，则应该从收集的响应中将其排除；

– 如果一个评价者对> 15%测试项的中距锚点评分高于90分，则应该从收集的响应中将其排除。如果>25%的评价者对中距锚点评分高于90分，则可能表示该测试项没有明显地被此锚点处理劣化。在此情况下，不应根据对该项的评分将评价者排除。

如果需要，可以在所有评价者已经完成他们的测试之前执行这个初步阶段(使测试实验室在测试完成之前评估他们是否有足够数量的可靠评价者)。

可能有益的是研究数据来确定不正确的异常数据点，以便对其进行进一步的分析。一个合适的方法是采用将指定给一个特定测试条件j和音频序列k的各个等级与所有等级的四分位值间距相比较。

中值$\hat{x}$和四分位值Q应该如下计算：

$$\hat{x} \coloneqq Q_2(x_{jk}) = \text{median}(x) \coloneqq \begin{cases} x_{jk\frac{n+1}{2}}, n \text{ odd} \\ \frac{1}{2}\left(x_{jk\frac{n}{2}} + x_{jk\frac{n}{2}+1} \right), neven \end{cases}$$，x 按大小增序排序，且

$$Q_1(x_{jk}) = \begin{cases} \text{median}\left(x_{jk1}, \ldots, x_{jk\frac{n+1}{2}}\right), & n \text{ odd} \\ \text{median}\left(x_{jk1}, \ldots, x_{jk\frac{n}{2}}\right), & n \text{ even} \end{cases},$$

$$Q_3(x_{jk}) = \begin{cases} \text{median}\left(x_{jk1}, \ldots, x_{jk\frac{n+1}{2}}\right), & n \text{ odd} \\ \text{median}\left(x_{jk\frac{n}{2}+1}, \ldots, x_{jkn}\right), & n \text{ even} \end{cases}.$$

四分位值间距计算公式为$\text{IQR}(x) := Q_2(x) - Q_2(x)$。

在本文中，异常值属于集$O(x_j\text{k})$：

$$O(x_{jk}) := \left\{x_{jk} \mid x_{jk} > Q_3(x_{jk}) + 1.5 \cdot IQR(x_{jk})\right\} \cup \left\{x_{jk} \mid x_{jk} < Q_1(x_{jk}) - 1.5 \cdot IQR(x_{jk})\right\}.$$

如果一个受测者给予一个特殊激励的等级x和被测系统是O(x)的元素，则应该检验该分级的理由。对一个测试进程记录的检验可以暴露出设备的技术问题或人为错误。询问评价者可以暴露出所给出的等级是否真的代表了他们的主观观点。如果存在异常数据点的理由显示为是一个错误，则可以在最后分析之前将其从数据集中去除，并在测试报告中注明去除它们的理由。

应用后筛选方法可以明确测试结果中的趋势。但是，应该牢记，评价者对不同伪像的敏感性是可变的，应该小心谨慎。通过增加聆听评价组的人数，将会减少任何个别评价者评级的影响。

4.2 评价组的人数

如果可以估计由不同评价者所给出评级的方差，且实验要求的解析度已知，就可以确定一个聆听评价组的适当人数。

在一个聆听测试条件在技术和表现上都受到严格控制的情况下，经验显示来自不超过20位评价者的数据经常足以从测试得出适当的结论。如果分析可以作为测试执行的继续，当已经达到从测试中得出适当结论的适当统计显著性水平时，则不需要对更多评价者进行处理。

如果出于任何原因而不能实现严格的实验控制，则可能需要更大的评价者数量来得到要求的解析度。

一个聆听评价组的人数不仅仅是出于对期望解析度的考虑。原则上，来自在本建议书中所涉及实验类型的结果仅仅对实际参与到该测试中的有经验收听者组有效。因此，通过增加聆听评价组的人数，可以宣称结果是对相当一般的一组有经验收听者进行的，并且因此有时可以认为更令人信服。也可能需要增加聆听评价组的人数来容忍评价者对不同伪像灵敏度有所变化的可能性。

5 测试方法

MUSHRA测试方法采用具有与参考信号相同的完整带宽的原始未处理节目资料(它也被用作一个隐藏参考)以及许多强制性隐藏锚点。

可以使用附加的隐藏锚点，最好是其他相关ITU-R建议书的那些测试项。因为锚点的特性可以对一个测试的结果有明显的影响，一个非标准锚点的设计应该考虑附件5中描述的最佳锚点特性。应该在一个测试报告中详细描述该测试所使用的任何非标准锚点的属性。

5.1 测试信号描述

建议序列的最大长度应该大约为10 s，最好不超过12 s。这是为了避免收听者疲劳，增加收听者响应的抗干扰性和稳定性，和减少聆听测试的总时长。这个时长还必须要让内容在整个信号时长上一致，这将增加收听者响应的一致性。此外，一个较短的时长也将使收听者能够比较测试信号的一个较大连续部分。

如果信号太长，收听者响应受测试信号最初和最后效应或者隔离环形区域的驱动，它们可能在整个测试信号时长上在频谱和时间特性上极大地变化。缩短测试信号时长的目的是减少这种可变性。但是，这种限制可能不适合于某些场景。一个实例可能是涉及一个声音长且缓慢移动轨迹的测试。在决定必须使用一个较长激励的这些受限情况中，必须要在最终测试报告中记录要求增加时长的理由。

处理的信号集包括所有被测信号和至少二个附加“锚点”信号。标准锚点是一个截止频率为3.5 kHz的原始信号的低通滤波版；中级质量锚点具有一个7 kHz的截止频率。

锚点的带宽对应于控制电路的建议书（3.5 kHz），用于分别按照ITU-T G.711、G.712、G.722和J.21建议书对广播、遥测电路(7 kHz)和临时电路(10 kHz) 中监视和协调的目的。

3.5 kHz低通滤波器的特性应该如下：

f_c = 3.5 kHz

最大通带波动= ±0.1 dB

在4 kHz处的最小衰减= 25 dB

在4.5 kHz处的最小衰减= 50 dB。

附加锚点目的是要提供被测系统如何与已知音频质量级别比较的指示，并且不应该用于在不同测试之间重新度量结果。

5.2 培训阶段

为了得到可靠的结果，在测试之前的特殊培训阶段培训评价者是必须的。已经发现这个培训对得到可靠结果非常重要。培训至少应该使受测者面对测试期间将会经历的损伤和所有测试信号的全部范围和特性。这可以通过采用多种方法来实现：一个简单的磁带回放系统或互动计算机控制系统。在附件1中给出了指导。培训还应该被用来保证评价者熟悉主观测试的装置 (例如测试软件)。

5.3 呈现激励

MUSHRA是一种具有隐藏参考和隐藏锚点的双盲多激励测试方法，然而，ITU-R BS.1116建议书采用一种“带有隐藏参考的双盲三激励”测试方法。MUSHRA方式被认为更适合于评估中等和大的损伤[4]。

在一个涉及小损伤的测试中，对于受测者，困难的任务是检测可能出现在信号中的任何伪像。在此情况下，为了使实验者能够评估评价者成功检测这些伪像的能力，测试中必须要有一个隐藏的参考信号。相反，在一个中等和大损失的测试中，受测者在检测伪像中没有困难，并且因此不需要隐藏参考用于此目的。然而，当受测者必须对各种伪像的相对烦扰分级时，出现了困难。此时，受测者必须权衡其对一种类型伪像相对于一些其他类型伪像的偏好程度。

使用一个高质量参考引入了一个令人关注的问题。因为新的方法要被用于评估中等和大的损伤，预计参考信号相对于测试项的感觉差别会相对大。相反，属于不同系统的测试项目之间的感觉差别可能会非常小。其结果是，如果使用一个多次尝试测试法(如在ITU-R BS.1116建议书中所使用的)，评价者可能非常难于在各种损伤信号之间进行准确区分。例如，在一个直接配对比较测试中，评价者可能认为系统A好于系统B。但是，在每个系统仅仅与参考进行比较的情况下(即，不直接相互比较系统A和系统B)，可能失去这二个系统之间的差别。

为了克服这个困难，在MUSHRA测试方法中，受测者可以根据意愿在参考信号和任何被测系统之间切换，通常采用一个计算机控制的重放系统，尽管也可以利用使用多CD或磁带机的其他装置。受测者被给予一系列试验。在每个试验中，给予受测者带有参考版本、低等和中等锚点、以及被测系统所处理的测试信号的所有版本。例如，如果一个测试包含8个音频系统，则允许受测者在11个测试信号和开放参考之间近乎即时地切换(1个参考 + 8个测试系统 + 1个隐藏参考 + 1个隐藏低等锚点 + 1个隐藏中等锚点)。

因为受测者可以直接比较损伤信号，此方法提供了一个完全配对比较测试的优点，受测者在其中可以更容易地检测损伤信号之间的差别，并对它们对应地分级。这个功能允许在指定给系统的分级中有一个高度的解析度。但是，重要的是要注意到，评价者将通过将该系统与参考信号以及与每个试验中的其他信号进行比较来得出他们的等级。

建议在任何试验中应包括不超过12个信号(例如，9个被测系统、1个隐藏低等锚点，1个隐藏中等锚点和1个隐藏参考)。

在不常发生的要比较大量信号的情况中，可能需要一个实验分块设计，应该对其进行详细报告。

在ITU-R BS.1116建议书的测试中，评价者倾向于通过以一个检测程序开始，跟随一个分级过程，来处理一个给定的试验。从按照MUSHRA方法进行测试得出的经验显示，评价者倾向于以一个粗略的质量估计开始一个进程。后面跟随着一个分类或排行处理。在这之后，该受测者执行分级处理。因为排行是以一种直接的方式完成，中级音频质量的结果很可能比使用ITU-R BS.1116建议书的方法更加一致和可靠。此外，最小循环时长是500 ms，并且应该将一个5-ms升余弦包络的渐入和渐出应用于所有循环的内容。在测试系统之间切换的

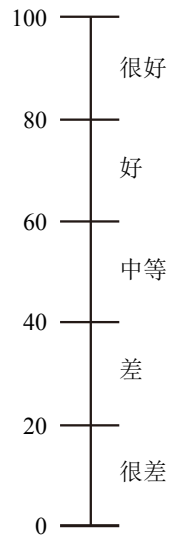
所有内容应该包括一个升余弦包络的5 ms渐入和5 ms渐出。在任何测试期间，当在测试系统之间过渡时，任何时候都不应该使用一个交叉渐变。这些改变目的是要在突然过渡比较期间减少使用频谱色度中的变化来确定测试信号并进行评分。

5.4 分级处理

要求评价者按照连续质量量表(CQS) 对激励进行评分。CQS由相同的图形量表构成(通常是10 cm或更长)，它被分成5个相等的间隔，带有如图1中从上到下所示的形容词。

这个量表也被用于对图像质量的评估 (ITU-R BT.500建议书 – 电视图像质量的主观评价方法)。

图 1

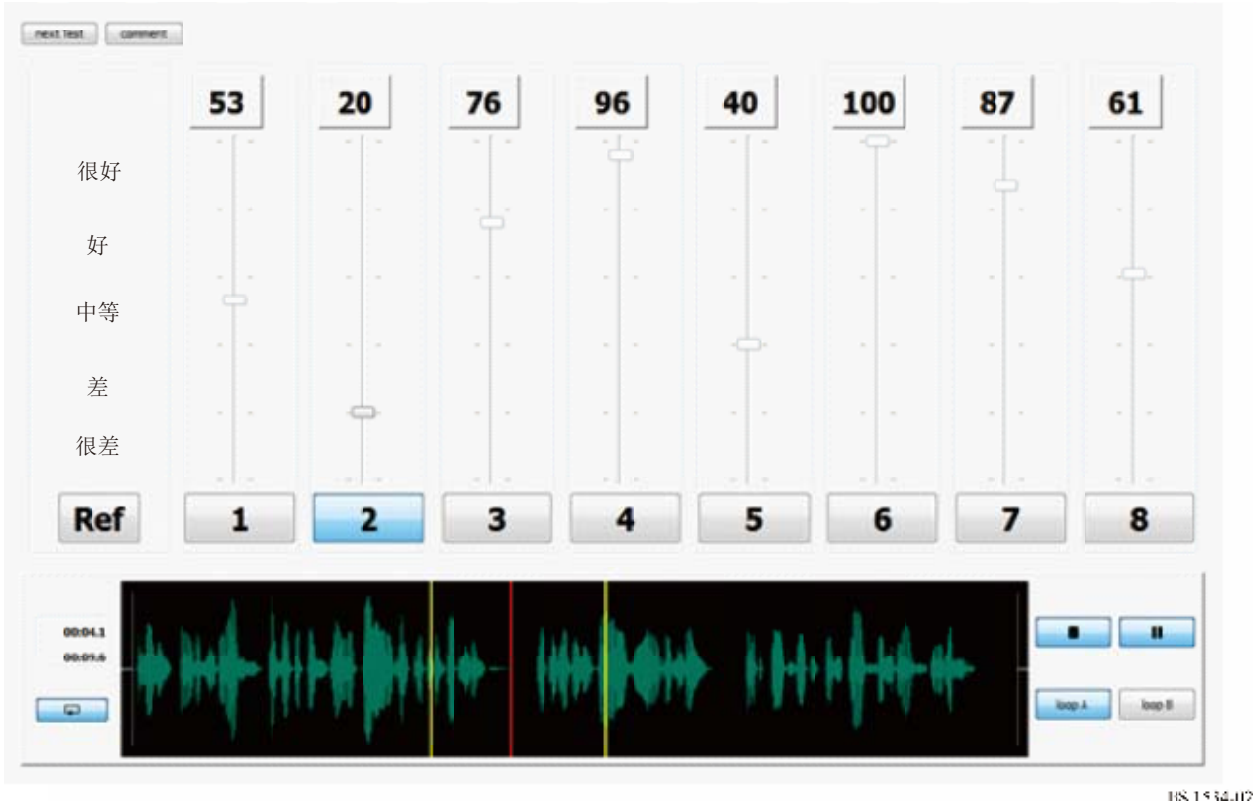


BS.1534-01

收听者以一种合适的形式记录其对质量的评价，例如，在一个电子显示器上采用一个游标(见图 2)，或者使用一支笔和纸质量表。应采用一个类似于图2中所显示的装置，将限制受测者只能调整分配给他或她当前正在收听项的分数。在附件2中可以找到对界面设计的一些指导。要求评价者按照5间隔CQS对所有激励的质量进行评分。

图 2

一个用于MUSHRA测试的计算机显示的实例



与ITU-R BS.1116建议书相比，MUSHRA方法具有同时显示很多激励的优点，这样受测者就能够在它们之间直接进行任何比较。相比于采用ITU-R BS.1116建议书的方法，采用MUSHRA方法进行测试所需的时间可以被极大地减少。

5.5 记录测试进程

在处理指定分数时观察到一些异常的事件中，记录产生这些分数的事件是有用的。这样做的一种相对简单的方法是制作整个测试的视频录像和音频录音。在一组结果中发现一个异常等级的情况中，可以检查磁带录音来尝试确定原因是人为错误还是设备故障。

6 属性

以下所列举的是对单声道、立体声和多通道评估特定的属性。最好在每种情况中评估属性“基本音频质量”。实验者可以选择定义和评估其他属性。

在一个试验期间应该仅仅对一个属性进行评级。当要求评价者在每个试验中评价多于一个的属性时，他们可能因尝试对一个给定激励回答多个问题而变得负担过重或困惑，或者二者兼而有之。这可能会对所有问题产生不可靠的分级。如果要独立判断多个音频特性，建议首先评估基本音频质量。

6.1 单声道系统

基本音频质量：这个单一的普遍属性被用于判断在参考和对象之间检测到的任何和所有差别。

6.2 立体声系统

基本音频质量：这个单一的普遍属性被用于判断在参考和对象之间检测到的任何和所有差别。

以下附加的属性可能会被关注：

立体声像质量：这个属性与参考和对象之间在声像位置和感觉深度和音频事件真实感上的差别相关。尽管一些研究已经显示出立体声像质量可能受到损伤，但是还未进行足够的研究来表明是否能保证对立体声像质量进行有别于基本音频质量的分别评分。

注 1 – 直到1993年，大多数对立体声系统小损伤主观评估的研究一直唯一地使用属性基本音频质量。因此，属性立体声像质量作为在那些研究中的一个普遍属性隐含或者明确地包含在基本音频质量中。

6.3 多通道系统

基本音频质量：这个单一的普遍属性被用于判断在参考和对象之间检测到的任何和所有差别。

以下附加的属性可能会被关注：

前置图像质量：这个属性与前置声源的位置相关。它 包括了立体声像质量和清晰度损失。

环绕印象质量: 这个属性与空间印象、周围环境或特定方向性环绕效果相关。

7 测试素材

应该采用呈现期望应用的典型广播节目的关键素材，以便揭示被测系统之间的差别。只要素材突出了被测系统，它就是关键素材。没有能够被用来在所有情况下评价所有系统的普遍适用节目素材。相应的，必须对要在每个实验中测试的每个系统寻找关键的节目素材。对适当素材的搜索通常耗费时间；但是，除非为每个系统找到真正关键的素材，否则实验将无法揭示系统之间的差别，并且将是无结果的。一个小的专家收听者分组应该从可能候选项的大量选择中选出测试项。这个选择过程必须包括所有测试系统，并在测试总结中记录和报告。

必须要从实验和统计上说明未能找到系统之间的差别不是因为可能由于选择了不适当的音频素材或者实验的任何其他弱势方面而引起实验的不敏感。否则，不能认为这个“零”发现有效。

在搜寻关键素材中，应该允许能被认为是潜在广播素材的任何激励。不应包括被故意设计用来破坏一个特定系统的合成信号。一个节目序列的艺术性或知识性内容既不应该有吸引

力，也不应该有争议或令人厌烦，以至于受测者从专注于检测损伤上分心。应该考虑在实际广播中每种节目素材出现的预期频率。但是，应该明白，广播素材的特性有可能随着将来音乐风格和偏好的改变而改变。

当选择节目素材时，重要的是要精确定义要被被评价的属性。应该将选择素材的责任赋予对预期损伤具有基本了解的一组熟练评价者。他们的起始点应该基于范围非常广泛的素材。这个范围可以借助专用的录音来扩展。

出于准备正式主观测试的目的，每个片段的响度需要在将其录到测试介质上之前由熟练评价者分组主观地进行调整。这将允许对该测试介质之后的使用采用为在一个测试试验内所有节目项设置的固定增益。

对于所有测试序列，熟练评价者组应该聚集并统一到各个测试片段的相对声音水平上。此外，专家们必须对该序列作为一个整体统一到相对于校准水平的绝对重现声压水平。

根据EBU R.68建议书 (见ITU-R BS.1116建议书的§ 8.4.1)，可以将具有校准信号水平的一个猝发音(例如，1 kHz, 300 ms, –18 dBFS)包括在每个录音的开始，使其输出校准水平调整到重现信道所要求的输入校准水平。猝发音仅仅用于校准的目的：在测试期间它不应被重放。应该控制声音节目信号，使得峰值的幅度仅仅难得超过ITU-R BS.645建议书中所规定的最大允许信号的峰值幅度 (校准水平以上9 dB的一个正弦波)。

包括在一个测试中片段的可行数量是变化的：它应该对每个被测系统都相等。一个合理的估计是被测系统数量的1.5倍，最小值是5个片段。由于此任务的复杂性，被测系统对于实验者应该是可以用的。只有当定义了一个适当的时间安排表，才能够达到一个成功的选择。此外，由于在音频编解码当中使用时变比特率，建议对更长的序列进行编码，并且在聆听测试中使用每个序列的一部分。

应该采用一个参考下混音测试双通道回放条件下的一个多通道系统的性能。尽管在一些环境下使用一个固定的下混音可能会被认为具有限制性，但是毫无疑问，对于广播电台使用从长远上是最敏感的方案。参考下混音的等式 (见ITU-R BS.775建议书) 为：

$$L_0 = 1.00L + 0.71C + 0.71L_s$$

$$R_0 = 1.00R + 0.71C + 0.71R_s$$

对参考双通道下混音性能关键评估的合适测试片段进行预选应该基于双通道下混音节目素材的重现。

8 聆听条件

在ITU-R BS.1116建议书中规定了对包括多通道音响系统的音频系统中小损伤主观评价的方法。为了评估具有中级质量的音频系统，应该采用在ITU-R BS.1116建议书的§§ 7和8中所列出的聆听条件。

耳机或扬声器都可以用在测试中。在一个测试进程中不允许二者都使用：所有评价者必须使用相同类型的换能器。

对轮流馈入到每个重现通道输入(即，一个功率放大器及其关联的扬声器)的一个具有等于“校准信号电平”(按照ITU-R BS.645建议书是0 dBu0s；按照EBU R.68建议书是一个数字磁带录音削波电平以下–18 dB，) r.m.s. 电压的测试信号，应该调整放大器的增益来给出参考声压水平(IEC/A加权，缓慢)：

$$L_{ref} = 85 - 10 \log n \pm 0.25 \text{ dBA}$$

此处，n是总装置中重现通道的数量。

在一个进程中允许由受测者对聆听水平进行分别调整，并应该限制在相对于由ITU-R BS.1116建议书所规定参考水平±4 dB的范围之内。对一个测试中测试项之间的平衡应该由选择组以评价者通常不需要对每一项进行分别调整这样一种方式来提供。

不允许在一项之中进行水平调整。

9 统计分析

对每个测试条件的评价被从评分表上长度度量线性地变换到0至100范围中的归一化评分，其中，0对应于量表的底部 (很差质量)。然后如下计算绝对评分。

可以根据所满足的统计假设进行参量或非参量统计分析 (见§ 9.3.3)。对涉及参量统计分析的指导请见附件 4。

9.1 数据可视化和试探性数据分析

统计分析应该总是以原始数据的可视化开始。这可以结合用于正太分布的带有拟合曲线的直方图、箱形图、或分位图(QQ图)。

箱形图数据可视化将提供对数据描述性汇总中异常值存在和影响的显示。应该进行这个可视化来识别来自所有中级评价者各自评分的展开与偏差。应该进行一个直方图可视化来确定存在一个基础多模分布。如果一个多模分布在数据中被清晰地可视化，建议实验者分别分析这个分布。

要评价多模等级b，可以使用以下公式：

$$b = \frac{g^2 + 1}{k + \frac{3(n-1)^2}{(n-2)(n-3)}}$$

其中：

n: 样本量

g: 有限样本的偏斜度

k: 聆听测试结果的超峰度。

这个系数将处于0和1之间。较高的值(> 5/9)可以解释为是多模性的显示。

根据对这些图的视觉检查、b、和关于观察到样本的基础总体，应该可以决定是否可以假设已经观察到一个正太分布。如果拟合曲线明显地倾斜、直方图包含很多异常值、或者QQ图根本不是一条直线，就不应该认为样本是正太分布的。对后筛选之后仍保留下来的所有收听者归一化评分中值的计算将产生中值主观评分。中值应该如下计算：

$$\hat{x} = \text{median}(x) = \begin{cases} x_{\frac{n+1}{2}} & n \text{ odd} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & n \text{ even} \end{cases}$$

x按大小排序。

分析的第一步是对每个呈现计算中值评分$\overline{\eta}_{jk}$。按照η_{ijk}是对一个给定测试条件j和音频序列k的观测者i的中值评分，而$\hat{\eta}$是样本的中值(所有观测者，所有条件，所有音频序列)。

类似地，可以对每个测试条件和每个测试序列计算总的中值评分$\overline{\eta}_j$和$\overline{\eta}_k$。

尽管使用平均值对一些分析方法是必须的，像ANOVA (见§ 9.3)，计算中值是一个居中趋势替代方法。中值提供一个可靠的居中趋势方法，它对样本集很小、非正太分布、或数据集包含明显异常值的情况最佳。有可能有很多这些担心不太被保证的测试场景。但是，由于标准化测试最大优点之一是对评分进行跨用户和场景比较和解释，确定最可靠和对可能改变有效性或减少测试间转换因素最不敏感的分析方法是有益的。

以这种方式，可以应用非参量统计。当应用非参量数据分析时，应该从可用方法计算平均值和95%置信区间，例如采用一个普通自举算法。

可以采用平均绝对偏差计算对中值误差的度量：

$$\hat{\tau} = \Sigma |Y_i - \hat{\eta}| / n$$

四分位间距(IQR) 被建议作为对中值的一种置信度量。它是第一和第三个四分位之间的差别评分：$IQR = Q_3 - Q_1$。此公式在§ 4.1.2中给出。如果结果是正太分布，IQR表示二倍平均绝对偏差。

建议统计显著性应该认定为是95%的显著性水平。不规则分布的非参量测试是统计显著性的可靠度量。和参量统计分析不同，它们不做关于数据基础分布的假设，并且对很多与使用较小样本量相关联的担心不太敏感。

如果数据真的如虚假设下所假设是随机的，对不规则分布的一个可靠非参量测试 (排列检测)可以确定在二种测试条件之间将出现一个观测差的概率。在此测试中测量到的概率是从实际数据分布确定的一个真实度量，而不是对基础分布假设一个指定形状的推断度量[5]。这种形式的测试要求通用重新抽样技术，例如自举和Monte-Carlo模拟技术，由于现代计算速度的加快，它们现在是轻易可以实现的[6]。附件 3中给出了对此测试方法的进一步描述。

9.2 功效分析

如果作为先验分析来应用，功效分析可以有助于估计聆听测试所需要的样本量，而在一个后验分析中，有助于估计测试的功效或类型II误差。给定了效应量 $d = \frac{\bar{x}}{s}$、显著性水平a和统计功效$1 - \beta$，一个*先验*分析可提供实验所需的样本量。

相比之下，给定了效应量 $d = \frac{\bar{x}}{s}$、显著性水平a和样本量N，一个后验分析提供该测试的功效1-β 或类型II误差β。类型II误差β 是效应d存在于总体中但未被该测试认为显著的概率。例如，如果一个测试宣布质量不受该系统的影响，$1 - \beta$就是损伤被该测试证明的概率。[2]

9.3 ANOVA的应用和使用

9.3.1 引言

本节集中于采用方差分析(ANOVA)进行参量统计的要求。由于ANOVA模型的强健性(见[7] [8] [12] [13])及其统计功效[3]，它对使用ITU-R BS.1534建议书中方法采集的数据是一个非常适合的方法。因为ANOVA F统计对非正太数据分布和方差的不均一性都相当强健，假设测试专注于误差或残留的特性。

对与参量统计相关的一般假设的进一步阅读请参见附件 4。

9.3.2 一个模型的规范

强烈建议，在实验设计期间(见§ 3)，在独立变量(例如，SAMPLE、SYSTEM、CONDITION等)和相关变量(例如，基本音频质量或收听尝试等)方面对模型进行了全面规定。还应该在模型规范阶段规定每个独立变量的水平。

当定义一个分析模型时(例如，采用方差分析ANOVA或重复测量ANOVA)，重要的是要包括所有重要变量。遗漏重要变量，例如独立因素的2或3方相互作用，可能导致该模型的错误规范，这可能进一步导致数据分析的差的解释方差(R^2)和潜在误解。

9.3.3 参量统计分析检验表

这个检验表是提供作为对数据审查、基本假设测试(参量和非参量)以及参量统计基本步骤的一个简短指导。这个检验表的重点是在方差分析的要求上，作为对来自ITU-R BS.1534建议书实验的数据进行分析的一个适当方法。要得到一个完整的指导，请读者参考关于统计学的教科书 (例如，[8] [11] [9])。

---

[2] 存在有很多工具，例如G*Power [16]，用于对未知总体的分布自动执行功效分析，而对未知总体估算功效更难。

[3] 通常建议选择该数据所允许的最强大的统计分析方法[9] [10]。

- 试探性统计学[4]
  - 审查数据结构是否正确且如预期
  - 检查缺失数据
  - 研究数据分布的正态性
  - 审查其他潜在的数据分布(单峰、双峰、倾斜等)
- 单维性
  - 审查该量表是否被评价者共同使用[5]
  - 测试该数据在特性上是否是单维的
  - 主要分量分析、Tucker-1图、或Cronbach的alpha
- 观测的独立性
  - 这通常是在实验方法中定义，并且不容易被统计地测出来。应该确保数据是独立收集的，即，通过采用双盲实验技术，并且确保评价者互相不影响。
- 方差均一性[6]
  - 测试每个独立变量展示相似方差的假设。
    - 对独立变量的每个水平采用并排箱形图的视觉观察；作为经验法则，均一性可能最大以一个4倍的因子变化
    - Brown和Forsythe的测试或Levene统计可以被用来评估方差的均一性
- 残值的正太分布
  - 测试残值的正太分布
    - Kolmogorov-Smirnov D测试或K-S Lillefors测试或Levene测试
    - 正太概率图(有时被称为P-P图)或分位图(经常被称为QQ图)也能够被用作对正太分布的视觉测试
- 异常值检测
  - 当被证实后，应该筛选出或者可以去除异常值。对这个问题的指导在§ 4.1.2中提供。
- 分析
  - 方差分析(ANOVA) – 通用线性模型或重复测量ANOVA模型
    - 使用一个合适的ANOVA模型，例如，通用线性模型(GLM)或重复测量ANOVA模型：更多细节在附件 4中提供
    - 按照实验设计详细说明模型

---

4 这同等地应用于参量和非参量统计。

5 已经在部分总体对涉及特殊伪像评估有不同观点的情况下观察到多维性。

6 对应用ANOVA要求，但对rmANOVA不要求 (见附件 4)。

- 在可能的情况下包括2和3方相互作用

– 采用该模型和结果分析数据
  - 审查用于描述相关变量的模型解释方差 (R^2)
  - 审查残值误差分布
  - 审查重要和非重要因素
  – 可以对模型进行迭代，以消除异常值和非重要因素
– 事后分析测试
  – 应用事后分析测试确立在ANOVA中相关因素（或因素相互作用）明显情况下平均值之间差别的明显性。
  – 有很多不同的事后分析测试可用，具有不同的辨识水平，例如，Fisher的最小显著差异法(LSD)、Tukey的真实显著差异法(HSD)等。
  – 建议与显著性水平一起报告效应量。
– 得出结论
  – 一旦已经进行了分析，通过对原始或ANOVA建模数据(有时称为估算边缘均值)绘制平均值和相关的95%置信区间来总结结果。
  – 在发现因素相互作用（例如，2或3方）显著的情况下，这些应该被绘制成图，来提供数据的概观。在这种情况下，仅仅绘出主要效应将会提供对数据的一个概观，带有混淆的相互作用效应。

在附件4和在普通统计和应用的文字中可以找到对ANOVA模型使用的进一步指导，例如，[11] [13] [15]。

10 测试报告和结果呈现

10.1 概述

结果的呈现应该以一个用户友好的方式来进行，这样，不论是缺乏经验或是专家的任何读者都能够获得相关信息。读者首先希望看到总体实验结果，最好是图形的形式。可以采用更详细的量化信息支持这样的呈现，尽管充分详细的数字分析应该是在附件中。

10.2 测试报告的内容

测试报告应该尽可能清晰地传达该研究的基本原理、使用的方法和得出的结论。应该呈现足够的详细信息，使有见识的人大体上能够复制该研究，以便根据经验对结果进行检验。但是，该报告没有必要包含所有各个结果。一个见多识广的读者应该能够理解并对该测试的主要细节做出评论，例如该研究的基本原因、实验的设计方法与执行、和分析与结论。

应该对以下问题予以特别注意：

– 对结果的一个图形呈现；

– 对所选有经验评价者的筛选和规范的一个图形呈现；
– 实验设计的定义；
– 测试素材的规范和选择；
– 有关用来处理实验素材的系统的一般信息；
– 测试配置的详细信息；
– 聆听环境和设备的详细物理信息，包括房间尺寸和声学特性、换能器的类型和放置、设备电气规范 (见注 1)；
– 实验设计、培训、介绍、实验序列、测试步骤、数据生成；
– 数据的处理，包括描述性和分析推论统计的详细信息；
– 在测试中使用了锚点；
– 在结果分析中使用了哪种后筛选方法– 这将包括排除异常值或未经培训收听者的方法；
– 测试是否采用ITU-R BS.1534建议书或ITU-R BS.1534-1建议书来完成；应该在文件中采用对所使用锚点条件的描述来清晰地给予指明；
– 让一个新用户能够产生用在测试中的任何锚点所必需而又未在本ITU-R BS.1534-2建议书中明确描述的适当定义和生成代码；
– 得出的所有结论的详细基本信息。

注 1 – 因为有一些证据表明，聆听条件可能会影响主观评价的结果，例如采用扬声器对比采用耳机重现，要求实验者明确地报告聆听条件，以及在实验使用的重现设备类型。如果计划采用不同换能器类型的组合统计分析，必须要检查这样一个结果组合是否可能。

10.3 结果呈现

对每个测试参数，必须给出评价等级统计分布的中值和IQR。

结果必须要和以下信息一起给出：

– 对测试素材的描述；
– 评价者数量；
– 结果的图形呈现；除了平均值和95%置信区间的呈现外，还应包括显示IQR的箱形图；应该报告被测系统之间的明显差别以及所应用的统计分析方法。

此外，当在箱形图视觉化之后数据支持这样的呈现时，还可以以适当的形式呈现结果，例如平均值和置信区间。

10.4 绝对等级

被测系统绝对平均等级、隐藏参考、和锚点的呈现对结果给出了很好的概览。但是，应该记住，这没有提供详细统计分析的任何信息。因此，观测不独立，而且仅仅是对绝对等级的统计分析而没有考虑所观测样本基础总体将不会得出有意义的信息。此外，应该报告按§ 9中建议所应用的统计方法。

10.5 显著性水平和置信区间

测试报告应该向读者提供关于所有主观数据固有统计特性的信息。应该陈述显著性水平，以及其他关于统计方法和结果的详细信息，它们将有助于读者理解。这样的详细信息可能包括置信区间或图形形式的误差条。

当然，没有“正确的”显著性水平。但是，习惯上选择数值0.05。原则上，根据所测试的假设，有可能采用单拖尾或双拖尾测试。

参考文献

[1] Stevens, S. S. (1951). Mathematics, measurement and psychophysics, in Stevens, S. S. (ed.), Handbook of experimental psychology, John Wiley & Sons, New York.

[2] EBU [2000a] MUSHRA – Method for Subjective Listening Tests of Intermediate Audio Quality. Draft EBU Recommendation, B/AIM 022 (Rev.8)/BMC 607rev, January.

[3] EBU [2000b] EBU Report on the subjective listening tests of some commercial internet audio codecs. Document BPN 029, June.

[4] Soulodre, G. A., & Lavoie, M. C. (1999, August). Subjective evaluation of large and small impairments in audio codecs. In *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*. Audio Engineering Society.

[5] Berry, K. J., Johnston, J. E., & Mielke, P. W. (2011). Permutation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(6), 527-542.

[6] Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. *Society of Industrial and Applied Mathematics CBMS-NSF Monographs*, 38.

[7] Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. Lawrence Erlbaum Associates, Inc.

[8] Keppel, G. and Wicken., T. D. (2004). Design and Analysis. *A Researcher's Handbook*, 4th edition. Pearson Prentice Hall.

[9] Garson, D. G. Testing statistical assumptions, Blue Book Series, Statistical Associates Publishing, 2012.

[10] Ellis, P. D. (2010). The essential guide to effect sizes. *Cambridge: Cambridge University Press*, *2010*, 3-173.

[11] Howell., D.C. (1997). Statistical methods for psychology, 4th Edition, Duxbury Press.

[12] Kirk., R.E., (1982). Experimental Design: Procedures for the Behavioural Sciences, 2nd edition. Brooks/Cole Publishing Company 1982.

[13] Bech, S., & Zacharov, N. (2007). Perceptual audio evaluation-Theory, method and application. John Wiley & Sons.

[14] Khan, A. and Rayner, G. D. (2003). Robustness to Non-Normality of Common Tests for the Many-Sample Location Problem, *Journal of Applied Mathematics & Decision Sciences*, 7(4), 187-206.

[15] ITU-T. Practical procedures for subjective testing, International Telecommunication Union, 2011.

[16] Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*,*(4)*, 1149-1160.

附件1的
附录 1
(规范性)

对评价者的指导

以下是应该给予或读给评价者听的指导类型的实例，以指导他们如何进行测试。

1 熟悉或培训阶段

聆听测试中的第一步是要熟悉测试过程。这个阶段被称为培训阶段，它是在正式评估阶段之前进行。

培训阶段的目的是使作为一个评估者的你达到以下二个目标：

– **A部分**：变得与所有被测声音片段和它们的质量水平范围相熟悉；和

– **B部分**：学会如何使用测试设备和分级量表。

在培训阶段的A部分，为了演示整个可能的质量范围，你将能够听到已经选定用于测试的所有声音片段。你将听到的声音项将多少是关键的，取决于比特率和所用的其他“条件”。图 3显示了用户界面。你可以点击不同的按钮来收听不同的声音片段，包括参考片段。以这种方式，你可以学会对不同的节目项欣赏一些不同的质量水平。根据通用条件对这些片段进行分组。在此案例中确定了三个这样的分组。每一个分组包括四个处理过的信号。

在培训阶段的B部分，你将学会使用可用的回放和评分设备，它们将被用于评估声音片段的质量。

在培训阶段的期间，作为个人，你应该能够学会以分级量表的形式表达可听见的损伤。在培训阶段的任何时间，你不应该与其他评价者讨论你对量表的个人诠释。但是，鼓励你向其他评价者解释伪像。

在培训阶段所给的等级在真实测试中将不被考虑。

2 盲分级阶段

盲分级阶段的目的是邀请你使用质量量表指定你的等级。你的等级应该反映你对呈现给你的每个声音片段质量水平的主观判断。每个试验将包含9个要被分级的信号。每一项大约长10 s。你应该通过点击各自的按钮来收听参考、锚点、和所有测试条件。你可以按任何顺序收听这些信号任意多次。

对每个信号使用滑动器来显示你对其质量的观点。当你对你给所有信号的评级满意时，你应该点击屏幕下方的“登记评分”键。

图 3

培训阶段A部分用户界面实例的显示图

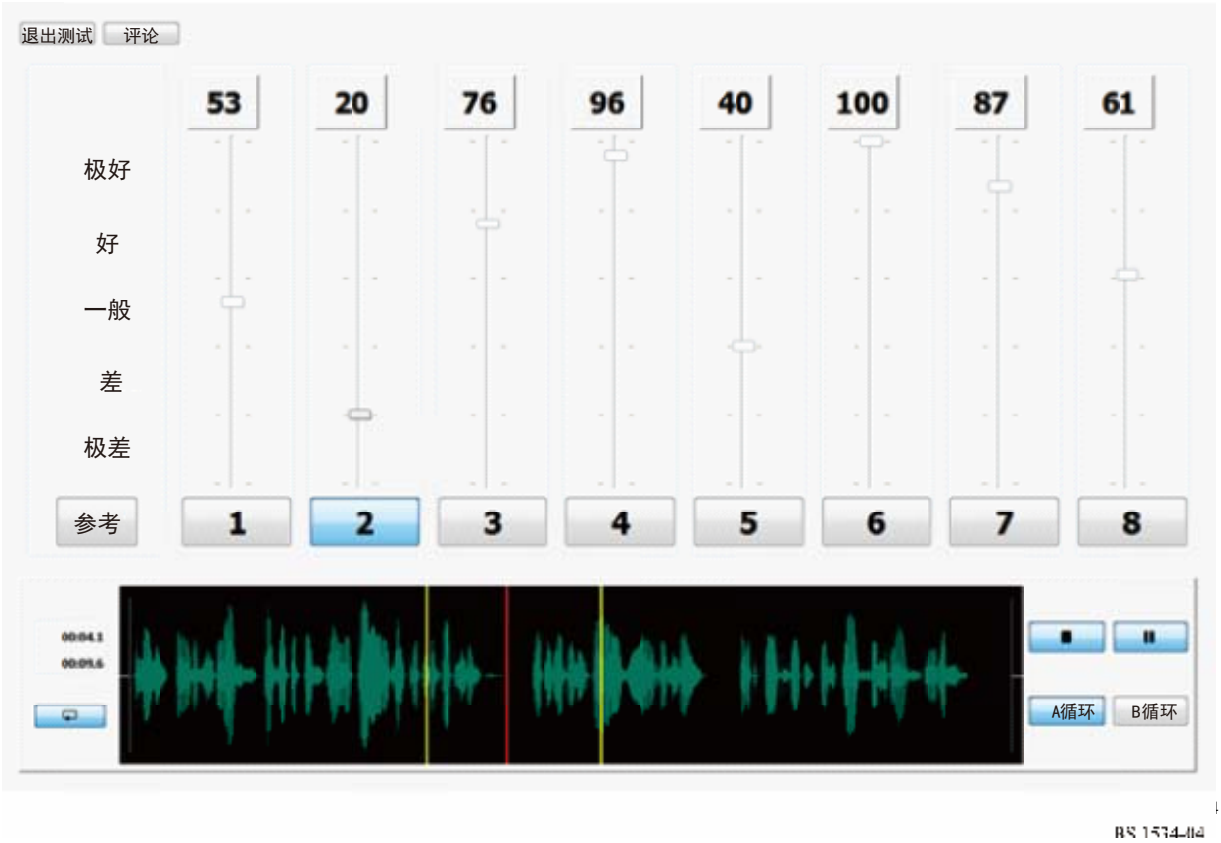
当指定你的等级时，你将使用如图1所给出的质量量表。

该分级量表连续从“极好”到“极差”。 一个0等级对应于“极差”分类的底部，而一个100等级对应于“极好”分类的顶部。

在评估声音片段中，请注意，你对测试中具有最低质量的声音片段不是必须应该给出一个“极差”分类中的等级。但是必须对一个或更多片段给予一个100等级，因为未经过处理的参考信号被包括作为要被分级的片段之一。

图 4

盲分级阶段中所用用户界面的实例



附件1的
附录 2
(参考性)

对用户界面设计的指导说明

对可能正在考虑下列问题的人提出了以下建议：

a) 按照MUSHRA方法制定执行主观测试的系统；

b) 执行这样的测试。

这些建议的目的是要增加测试结果的可靠性，并帮助对在处理测试评分期间可能会发现的任何不规则进行的分析。

用户界面的设计应该使一个受测者指定不符合它们真实意图的评分的可能最小。至此，应该采取步骤来确保从用户界面到在一个给定时间受测者正在收听的测试项经过处理的版本是清晰的。这可以通过仔细选择屏幕上指示符（例如，可点击的按钮）的颜色和亮度来解帮助避免一个受测者对一些颜色不敏感带来的潜在困扰。

应该确保受测者只能调整指定给当前正在收听项的评分。已经观察到，为了给他们听到的第一个而不是最后一个指定评分，一些评价者连续收听一个测试项的二个经过处理的版本。在这种情况下，有可能产生一个错误(特别是当屏幕上出现大量的控制时)，并且评分可

能会指定给并非想要的信号。为了试图减少这种可能性，建议在任一时刻允许使用的唯一控制是与当前正在收听信号相关联的那个，而指定给其他不是当前正在收听信号的控制应该被禁用。

附件1的
附录 3
(规范性)

对采用再抽样技术和Monte-Carlo仿真方法在二个样本之间进行非参量统计比较的描述

可以将随机化的非参量测试与普通再抽样技术一起使用，例如自举程序，来确定几乎任何统计结果的显著性。例如，可以用以下方式来计算在二个测试信号(样本数量 = *N*1和*N*2)之间所观测到的中值响应差的显著性：必须标明每个样本中值之间的实际差别，并且将其标记为$Diff_{ACT_1}$。然后将来自这些样本的所有数据汇聚成单独一个文件或向量。应使用一个自举程序，这样使得对每次迭代都采用*N*1和*N*2数量的抽取样本来排列汇聚组而不用更换。二个随机抽取样本中值之间的差别将被记录为$Diff_{EST_1}$。然后，可以重复这个程序10 000次，而且$Diff_{EST_N}$超过$Diff_{ACT_N}$的数量除以10 000的比值将产生一个对应的*p*值。如果$Diff_{EST_N}$超过$Diff_{ACT_N}$的总次数少于500 (500/10 000 = .05)，可以说这二个平均值之间差的显著性是.05级别，$p < .05$。

附件1的
附录 4
(参考性)

对参量统计分析的指导说明

1 引言

在§ 9中给出了对MUSHRA测试中结果基本参量统计分析的一个描述。但是，特别是当很多条件要相互比较时，像ANOVA这样的一个多项测试更适合于多对比较。本附件描述了这可以怎样完成。它包括分析的前提条件，并指出了当不满足这些时的替换选择。

MUSHRA测试采用一个*重复测试或受测者内设计*(可以在2004年版的Maxwell & Delaney中找到对这些概念的最好介绍)，其中二个受测者内因素(条件和音频素材)是完全交叉的，并且从收听者、音频素材和条件的每个组合至少得到一个评分。还可以有这样的情况，相同的

音频素材和条件的组合呈现给二个或更多不同的评价者分组，例如在不同的实验室中。在此情况下，有一个在分析中必须要考虑的附加受测者间因素*组*。

为了概括在一个和所有收听者总体相比要小的收听者样本中得到的结果，推论统计是必要的。例如，如果在聆听测试中评分显示在一个新编码器和一个已有编码器感觉音频质量之间有差别，则重要的是要回答一个问题，即，如果一个完全不同的收听者分组对这二个系统的音频质量进行评分，是否也能期望这个差别。关于MUSHRA聆听测试的特定设计，至少有一个人可能希望回答的三个问题(或者，从统计上来说，一个人希望测试的假设)，而此处描述的推论统计提供有效的答案。首先，主要关心的问题通常将会是被测系统之间(例如，参考和三个不同的编码器)感觉的音频质量差别是否会有所不同。其次，如果在聆听测试中音频系统是采用不同的测试素材来评估的，音频质量的评分是否取决于音频素材？第三，音频系统对感觉音频质量的影响是否在测试素材之间不同？回答这些问题的恰当方法首先是通过进行一个方差分析(ANOVA)来得到条件(音频系统)的主要效果、音频素材的主要效果、和条件 × 音频素材相互作用的显著性测试。当音频系统感觉质量之间的差别取决于音频素材时，就存在一个相互作用。请注意，由于潜在的相互作用，不建议对每个音频系统横跨音频素材汇聚评分，即使一个人对音频素材的效果或相互作用的效果不是特别感兴趣。则可以采用附加的比较来测试更多的特殊假设，例如，关于一对音频系统之间的感觉差别。

只要多于二个实验条件要进行比较，例如4个不同的编码器，则不适合将推论统计建立在多对比较之上。例如，如果$K = 5$ 个音频系统包括在测试中(4个编码器加参考)，则有$\binom{K}{2} = K(K-1)/2 = 10$ 对条件。在一个.05 α水平采用10对样本*t*测试对这10对的每一对的差别进行测试将导致所谓的族系类型I误差率膨胀。对每个分别的*t*测试，错误地拒绝二个编码器感觉音频质量之间没有差别的零假设的概率是α。

横跨C个这样的测试，产生至少一个类型I错误的概率是$1 - (1 - \alpha)^C$，对如我们所用实例的$C = 10$是0.40，而且因此比所期望0.05的α水平高很多。族系误差率可以通过对多重测试应用适当的修正来控制，例如稍后描述的Bonferroni修正或Hochberg (1988)程序。但是，带修正的按对*t*测试仍揭示了相关信息，部分是因为对所有平均值对进行的多重*t*测试使用了冗余信息 (每个平均值出现在多个测试中)。按对测试方式通常将不如采用适当的多项测试作用大(即，在检测条件之间差别上敏感性差)，对于MUSHRA测试，多项测试是一个方差重复测量分析(rmANOVA)。在后面，对不包含受测者之间因素的一个MUSHRA测试情况提供了一个数据分析的逐步描述。换句话说，假设仅仅对一组评价者进行测试，并且所有条件和音频素材的组合呈现给每个评价者至少一次。稍后将描述一个具有多于一个分组的设计的扩展(例如，当测试是在二个实验室中进行时)。

2 正态性测试

考虑一个相对于对统计测试有效性的响应测量正态性潜在偏差的影响是审慎的。对于每个评价者仅仅在一个实验条件中进行测试的受测者间设计，在一般线性模型的框架中进行的ANOVA就响应测量非正态性而言是惊人的强健(例如，[11]; [13]; [25]; [35])。

对如MUSHRA测试中的一个重复测量设计，我们首先注意到一个对在总体中感觉音频质量对所有条件都相同的零假设进行测试的替代方法。这等效于计算$K-1$正交对比，例如，通过构建K个条件间的差别变量，然后测试所有这些差别变量的总体平均等于0这个假设。例如，如果测试包含参考和二个编码器，则通过对每个受测者计算参考评分与编码器A评分之间的差别（$D1$）和编码器A评分与编码器B评分之间的差别(D_2)，可以创建二个差别变量D_1和D_2。重复测量ANOVA方式都假设这些差别变量是多重正太分布的。不幸的是，不像对于受测者间设计，非正太可能导致过于保守或过于宽松的类型I误差率 ([5]; [22]; [30]; [39])。这意味着，对于一个给定的α水平(例如，$\alpha = 0.05$)，尽管对所有条件相等平均值的零假设为真，ANOVA仍产生一个显著p值($p < \alpha$)的情况的比例将小于或高于正太值α。同样不像对于受测者间设计，简单地增加样本量解决不了这个问题[30]。有不断积累的证据表明，从峰态意义上，偏离对称比对正太分布的偏差具有严重得多的影响 ([4]; [18])。偏离对称的程度可以用分布的*偏斜度*来表示，它是三级标准化动差[8]。对于正太分布这样的对称分布， 偏斜度是0。*峰度*是对于平均值的标准化四级总体动差，并且描述了峰度和尾重 (图示见[9])。以前的模拟研究表明，对于小的偏离对称，rmANOVA仍将控制类型I误差率。但是，当前的研究状态尚不能以公式表达关于可接受的相对正太性偏差程度的精确法则。因此，建议对多变量正态性进行测试，并且报告偏斜度和峰度的经验估算。

重要的是要注意，作为RMANOVA基础的一般线性模型并不假设原始响应 (即，MUSHRA测试中的评分)是正太分布的。 相反，该模型假设*误差*是正太分布的。因此，必须对该模型的*残值*计算正态性测试或偏斜度和峰度测量，而不是对原始数据。幸好，大多数统计软件能够为每个分析过的实验条件保存残值，在当前的案例中是每个音频系统和音频素材的组合。这将对每个实验条件提供一个残值矢量。在每个矢量中，每个值表示一个评价者。

有多种多变量正态性测试可用，例如，Royston所建议的多变量Shapiro-Wilk测试 [34]、基于多变量偏斜度和峰度的测试 [10]及其他方式 [14]。应用这类测试的宏软件可以在SPSS (http://www.columbia.edu/~ld208/normtest.sps) 和 SAS (http://support.sas.com/kb/24/983.html) 得到，而且其他软件包也非常类似。所有主流统计软件包提供偏斜度和峰度单变量估算，可以对每个音频系统和音频素材组合的残值分别计算。DeCarlo 的 SPSS 宏软件 [9] (http://www.columbia.edu/~ld208/normtest.sps)还计算多变量偏斜度和峰度 [26]。应该报告单变量或多变量偏斜度和峰度的估算，以及多变量正态性测试的结果。

如果多变量正太性测试不显著，或者如果所有多变量或单变量测试未显示出对一个正太分布所预期值的显著偏斜度或峰度偏差，则rmANOVA的假设被满足。

但是，如果任何测试显示出相对正态性的一个显著偏差，或者如果任何试验条件中的偏斜度超过一个0.5的值 (作为一个初步的经验法则)，则产生了这些发现的结果应该是什么的问题。有二个一般问题，并且二者都与所讨论的缺少关于rmANOVA相对于正态性可接受偏差的法则相关。首先，多变量正态性测试是相当敏感，并且将经常检测到相对于正态性的负偏差。它们还将不仅检测到残值分布中的一个非对称性，而且分布的峰度或其他方面也在起着作用，然而非常可能仅仅非对称性导致了rmANOVA中非强健的类型I误差率。其次，如果对多变量偏斜度和峰度的测量是从该数据估算出来的[26]，这个信息不允许决定是否可以应用rmANOVA，同样是因为缺少关于相对于正态性可接受偏差的法则。这强调了报告偏斜度和峰度测量以及测试结果的需要。只要关于相对于正态性可接受偏差的有效法则变得可用，则可以采用改善的信息重新估算rmANOVA测试结果。如果相对于正态性的偏差看上去严重，例如，偏斜度估算显示高于1.0 [29]，则可以考虑rmANOVA的非参量替代方法，例如，采用重新抽样技术的测试或Friedman测试。但是，仍不清楚在什么情形中重新抽样技术能够解决非正态性的问题 [38]。Friedman测试不假设多变量正态性，而是假设方差对所有实验条件都相同 [36]，对实验数据并不经常是这样的情况。此外，Friedman测试是一个单变量测试。因此，即使等方差的假设成立，Friedman测试可以被用来检测在音频系统之间进行平均的效果，但它不能用来分析音频系统 × 音频素材的相互作用。

3 选择rmANOVA方式

对来自重复测量设计的数据，存在有很多用于测试受测者内部和之间因素效应的不同方式[21]。因为我们当前考虑的是一个不包含受测者（分组）之间因素设计的情况，并且因为我们假设没有丢失数据 (即，对每个收听者、音频素材、和条件组合的有一个有效评分)，有二种可以推荐的方式。当数据是多变量正太时，二者都提供对假想的有效测试，但是它们的统计功效可能不同 (即，检测相对于零假设偏差的灵敏度)，除其他因素外，取决于样本量。

这二个分析变化形式是 (a)具有对自由度的Huynh-Feldt修正的单变量方式，和(b) 多变量方式。可以在在其他地方找到这些方式的详细描述 [21]; [28]。在主流统计软件包中都可以得到这二种变化形式 (例如，R、SAS、SPSS、Statistica)。

由于数据的重复测量结构，在不同的条件和音频素材组合中得到的评分是相关的。例如，如果收听者给低质量锚点赋予一个高得出奇的评分，则他或她对编码器的评分也可能倾向于高于其他评价者的评分。单变量方式假设数据的方差协方差结构是球面的，这等于说以下所描述的差变量都具有相同的方差 [16]; [33]。但是，这个假设实际上对所有经验数据组不成立[21]。为了解决这个问题，在按照F分布计算p值时对自由度应用了一个修正因子。为此，从此数据估算相对于球形度的偏离量。因为替代的Greenhouse-Geisser [12]修正因子倾向于产生保守的测试(例如，[17]; [30])，所以建议Huynh-Feldt修正因子，称为$\tilde{\varepsilon}$ [17]。当该数据是正太时，具有Huynh-Feldt修正的单变量方式产生有效的类型I误差率，甚至对极小的样本量 ($N = 3$)。所有主流统计软件包提供修正因子$\tilde{\varepsilon}$ 和被修正的p值。

多变量方式采用一个替代但是等效的零假设公式。例如，考虑到在总体中的零假设，感觉的音频质量对所有条件是相等的。这等效于计算$K-1$个正交对比，例如，通过形成K个条件之间的差变量，然后测试所有$K-1$个对比总体平均值的矢量μ等于零矢量$\mu = 0$的假设。例如，如果有参考和二个编码器，则可以通过对每个评价者计算参考评分与编码器A评分之间差别（D_1）和编码器A评分与编码器B评分之间差别（D_2）来产生二个差变量D_1和D_2。采用多变量方式的rmANOVA是基于这些差变量并使用一个假设$\mu = 0$的多变量测试。以这种方式，不需要关于方差协方差矩阵的假设。对于按照一个多变量正太分布的数据，此测试是准确的，但它要求至少与因素级别数量同样多的评价者。因此，例如对9个条件（8个编码器加参考）呈现给仅仅8个评价者，则不能使用它。

除很多其他因素外，这二种方式的相对效能取决于样本量和受测者内部因素的因素级数。按照Algina和Keselman (1997)，如果$\tilde{\varepsilon} > 0.85$且$N < K + 30$，一个简单的选择法则将是使用具有Huynh-Feldt修正的单变量方式，此处N是评价者数量，而K是受测者内部因素的最大级数。在剩余的情况中，应该使用多变量方式。请注意，如果实验是在不同的实验室中进行的，则N是参加到该研究中的评价者总数(例如，10个评价者在实验室A中和10个评价者在实验室B中，对应于$N = 20$)。

4 实施选定的rmANOVA和可选的后处理测试

在这一步中，采用rmANOVA变化形式进行对条件、音频素材、及它们之间相互作用效应的综合测试。为了计算rmANOVA，大多数软件包要求可以一个“每个评价者一行”的形式得到数据，例如，SAS、SPSS、和Statistica。因此，数据表必须包含每个评价者仅仅一行，而对所有条件和音频素材组合的评分呈现为列(“变量”)。

双因素rmANOVA提供关于三个效应的信息。

1) 条件的主要效应

对大多数情况，这将是主要关注的测试。如果ANOVA显示出一个显著的条件效应，则可以拒绝在总体中感觉的音频质量对所有条件(参考、编码器1至k)相等的零假设。换句话说，此测试显示出，在总体中，感觉的音频系统音频质量之间有差别。作为一个效应量的度量，不可能使用Cohen[6]的d或其类似量之一，因为d不是对多于二个平均值比较定义的。在ANOVA内容中，普遍报告对相关性强度的度量。这些度量提供关于由关注的效应所导致的数据中方差比例的信息。这是与作为决定系数R^2基础相同的原理。大多数统计软件包可以计算部分η^2，按由该效应所引起的方差与效应方差和误差（残值）方差的和之比来计算。可以在Olejnik和Algina [31]中找到对相关性强度替代度量的讨论。

在对一个主要效应的显著测试结果之后，经常是关心确定此效应的起因。这可以通过计算特定的对比来实现。例如，一个人可能会关心一个新编码器的声音质量是否不同于三个已有系统的声音质量。为了回答这个问题，一个人可能首先对每个评价者计算三个已有编码器的平均评分，在音频素材间进行平均。作为结果，对每个评价者，将有(a) 对一个新编码器的评分，和 (b) 对三个其他编码器的一个平均评分。然后，这二个值与成对样本*t*测试相比较。请注意，因为该数据是来自一个重复测量设计，因此重要的是不要使用汇聚方差[27]。还要注意，这个对比也可能已经被作为一个计划的对比来测试，而不是执行ANOVA。通常建议使用显著性的双尾测试。但是，例如，如果有一个新编码器应该得到比已有编码器更好评分的*先验*假设，则将可能允许使用一个单尾拒绝区。

其他特殊对比可以使用相同的原理来计算。测试对比的一个更加一般的公式是计算一个在不同实验条件中所获得评分的线性组合，然后使用一个单样本*t*测试来决定此对比是否明显不等于0。 对于每个评价者i，可以计算一个对比值

$$\Psi_i = \sum_{j=1}^{a} c_j Y_{ij}, \quad \sum_{j=1}^{a} c_j = 0,$$

此处，Y_{ij}是评价者i在条件j中提供的评分 (对音频素材平均)，a是在此对比中考虑的条件数量，而c_j是系数。对以上实例，如果新编码器对应于$j = 1$，而三个其他编码器对应于$j = 2…4$，则选择$c_1 = -1$和$c_2 = c_3 = c_4 = 1/3$将提供对新编码器不同于其他三个编码器音频质量的假设的测试。

如果计算多于一个后验对比，则如以上所讨论，这将引入多重测试的问题。为了解决这个问题，建议应用Hochberg的[15] 随后可接受设置的Bonferroni程序。这个程序控制族系类型I误差率，而它比很多替代程序功能更强 [20]。在Hochberg程序中，首先计算m个关心的对比，并按p值排列它们。然后通过检验最大的p值来开始。如果此p值小于α，则所有假设被拒绝 (即，所有对比是显著的)。如果不是，则具有最大p值的t测试不是显著的，并且继续将下一个较小的p值与$\alpha/2$进行比较。如果较小，则此测试和所有具有较小p值的测试是显著的。如果不是，则具有第二大p值的测试不是显著的，并且继续将下一个较小的p值与$\alpha/3$进行比较。在更多正式项中，如果p_i （$i = m, m - 1, …1$）是按降序排列的p值，则对任何$i = m, m - 1, …, 1$，如果$p_i < \alpha/(m - i + 1)$，则$i' \leq i$的所有测试是显著的。

原则上，也有可能计算所有条件评分之间的后验成对比较。对一个重复测量设计，这将要求计算所有条件对之间的成对样本*t*测试。但是，不建议这种方式。考虑一个具有7个编码器和一个参考的实验。对这个8条件组，能够计算$8 \cdot 7/2 = 28$个成对测试，并且将不容易从这个大数量测试中得出有意义的信息。如果对所有成对差进行测试，则由于测试的数量大，应用Hochberg[15]程序来修正多重测试肯定是特别重要的。请注意，如果有据此进行成对样本*t*测试的相对于差别评分正态性偏差的证据，则非假设正态性的一个替代测试是符号测试。

应该注意，在一个显著主要效应之后，可能是没有后验对比或成对差别是显著的情况[28]，原因是rmANOVA和后验测试使用不同的统计信息。重要的是，rmANOVA是更合适的测试。因此，即使没有后验测试刚好是显著的，由ANOVA所显示的一个显著效应也保持有效。如果在一个显著综合测试 (ANOVA) 之后没有后验比较是显著的，则可以得出结论，这些音频系统在感觉声音质量上是不同的。也可以相互比较这些音频系统之间的差别。例如，对于显示出声音质量评分中最高差别的音频系统对，很可能这些成对差别将随着更大的样本量而变得显著。但是，必须得出结论，在当前的研究中，没有成对差别是显著的。

如果rmANOVA显示*没有*显著的主要条件效应，则这表示被测系统之间的差别很小。但是，由于有限的样本量，不能得出在总体中条件之间*没有*感觉音频质量差别的结论[3]。总体差别可以是零，或者考虑到样本量，效应量可能太小以至于不能检测到。如果进行了一个*先验*效应分析，即选定样本量足够以一个规定的概率检测到一个规定的效应量，则可以得出结论，该数据是针对预先规定量效应的证明。

这个发现可以被认为是编码器透明度的定义。如果没有进行一个先验功效分析，则，出于以上所解释的原因，推断编码器是透明的时候必须要小心谨慎。一个通常的后验近似解决方法是将p值与[0.2]相比较，而不是0.05。如果测试保持不显著，则这是一个比较强的在条件的感觉音频质量中缺少差别的指示。

2) 音频素材的主要效应

采用与以上相同的步骤和原理，对音频素材主要效应的测试提供关于根据测试素材评分系统性变化的信息。对大多数MUSHRA测试场景，这个效应并不应该是高度关注的，因为它与音频系统之间的差别不相关。

3) 条件与音频素材之间的相互作用

如果rmANOVA显示出条件和素材的一个显著相互作用，则音频系统对感觉音频质量的效应在测试素材之间是不同的。例如，对于一个高度压缩的流行歌曲，其中的编码伪像被素材中存在的失真成分掩盖了，参考和一个编码器可以评分相同。另一方面，编码器的声音质量评分可能会劣于一个三角钢琴高动态范围录音的参考。这个相互作用对一个MUSHRA测试通常将是关注的，因为它显示出音频系统之间的差别取决于测试素材。

在对相互作用效应进行一个显著综合测试之后，可以采用后验测试进一步地探察相互作用的特性。一种常见的方式是测试所谓的*简单主要效应*。例如，这些可以通过进行多个分别带有受测者内部因素条件的单因素rmANOVA来计算，对每个音频素材进行一次。这些分析将显示对哪个音频素材会有一个条件的显著效应。同样，Hochberg程序应该被用作对多重测试的一个修正。

如上所述，原则上可以采用分别的成对样本t测试和Hochberg程序来对所有条件和音频素材之间的成对差别进行测试。但是，成对比较的数量甚至将大于主要效应。例如，如果将8个音频系统与4个测试素材相组合，则有24个音频系统和测试素材的组合，对应于$24 \cdot 23/2 = 276$个成对测试。很明显，不能建议这种方式。

5 对包含一个受测者（分组）间变量设计的扩展

直到现在，我们考虑了一个没有受测者间因素的设计。当对不同的评价者分组进行测试时，例如在二个实验室中，或者音乐家对比非音乐家，应该进行哪个分析？

如果存在受测者间因素，则在所有分组中评价者数量是否相等（平衡设计）或在分组之间不同（非平衡设计）是至关重要的。

平衡设计。如果评价者数量对所有受测者间因素的所有级别相等，或者如果分组大小的差别不大于10%，则同样，用于自由度的具有Huynh-Feldt修正的单变量方式或多变量方式都可以用于进行rmANOVA [21]。该设计现在将包含受测者间因素条件和音频素材，并且至少一个受测者间因素(例如，实验室)。因此，rmANOVA将提供一个受测者间效应以及所有受测者内部和之间效应之间相互作用的附加测试。

例如，可能会得出条件 × 实验室相互作用是显著的，这将意味着音频系统感觉音频质量中差别对于实验室A不同于实验室B。请注意，这里我们假设对所有分组呈现完全相同的条件和音频素材组合。例如，如果不同的音频素材呈现在二个实验室中，则不能使用此处建议的方法。反之，将要求所谓的随机效应模型 [28]，它超出了本附件的范围。

非平衡设计。如果分组的大小相差大于10%，则不幸的是，单变量和多变量方式都不再提供有效的测试结果[21]。因此，强烈建议计划相等的分组大小，并因此避免这个问题。如果分组的大小不相等，则可以建议二个分析程序。第一个方式是改进的一般近似(IGA)测试[1]，而第二个方式是一个基于最大相似的混合模型分析的特殊变化形式[23]。IGA测试作为一个SAS宏是可行的。例如，该混合模型分析可以在SAS PROC MIXED中进行。对后一个分析，二种选择是重要的。首先，必须按照[19]的方法计算自由度，在SAS中这是通过在model陈述中设定ddfm=KR的选项来达到的。其次，必须对一个不均匀受测者间非结构化协方差结构（UN-H）进行拟合[23]，采用在重复陈述中的类型=UN组=groupingvar选项，此处，groupingvar是包含分组分类变量的名称。

参考文献

[1] Algina, J. (1997). Generalization of Improved General Approximation tests to split-plot designs with multiple between-subjects factors and/or multiple within-subjects factors. *British Journal of Mathematical and Statistical Psychology, 50,(2)*, 243-252.

[2] Algina, J., & Keselman, H. J. (1997). Detecting repeated measures effects with univariate and multivariate statistics. *Psychological Methods, 2*(2), 208-218.

[3] Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *British Medical Journal, 311*(7003), 485-485.

[4] Arnau, J., Bendayan, R., Blanca, M. J., & Bono, R. (2013). The effect of skewness and kurtosis on the robustness of linear mixed models. *Behavior Research Methods, 45*(3), 873-879. doi: 10.3758/s13428-012-0306-x.

[5] Berkovits, I., Hancock, G. R., & Nevitt, J. (2000). Bootstrap resampling approaches for repeated measure designs: relative robustness to sphericity and normality violations. *Educational and Psychological Measurement, 60*(6), 877-892.

[6] Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, N.J.: L. Erlbaum Associates.

[7] Conover, W. J. (1999). Practical nonparametric statistics (3rd ed.). New York: Wiley.

[8] Cramér, H. (1946). Mathematical methods of statistics. Princeton: Princeton University Press.

[9] DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2*(3), 292-307. doi: 10.1037//1082-989x.2.3.292.

[10] Doornik, J. A., & Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics, 70,(s1)*, 927-939. doi: 10.1111/j.1468-0084.2008.00537.x.

[11] Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237-288. doi: 10.3102/00346543042003237.

[12] Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*(2), 95-112.

[13] Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte-Carlo results in methodological research: The one-factor and two-factor fixed effects ANOVA cases. *Journal of Educational and Behavioral Statistics, 17*(4), 315-339. doi: 10.3102/10769986017004315.

[14] Henze, N., & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics-Theory and Methods, 19*(10), 3595-3617. doi: 10.1080/03610929008830400.

[15] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*(4), 800-802.

[16] Huynh, H., & Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact *F*-distributions. *Journal of the American Statistical Association, 65*(332), 1582-1589.

[17] Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational aand Behavioral Statistics, 1*(1), 69-82. doi: http://dx..org/10.2307/1164736.

[18] Jensen, D. R. (1982). Efficiency and robustness in the use of repeated measurements. *Biometrics, 38*(3), 813-825. doi: 10.2307/2530060.

[19] Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*(3), 983-997.

[20] Keselman, H. J. (1994). Stepwise and simultaneous multiple comparison procedures of repeated measures' means. *Journal of Educational and Behavioral Statistics, 19*(2), 127-162.

[21] Keselman, H. J., Algina, J., & Kowalchuk, R. K. (2001). The analysis of repeated measures designs: A review. *British Journal of Mathematical & Statistical Psychology, 54*, *(1)*, 1-20.

[22] Keselman, H. J., Kowalchuk, R. K., Algina, J., Lix, L. M., & Wilcox, R. R. (2000). Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping. *British Journal of Mathematical & Statistical Psychology, 53*,*(2)*, 175-191.

[23] Kowalchuk, R. K., Keselman, H. J., Algina, J., & Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted *F* tests. *Educational and Psychological Measurement, 64*(2), 224-242. doi: 10.1177/0013164403260196.

[24] Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models* (2nd ed.). Cary, N.C.: SAS Institute, Inc.

[25] Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research, 66*(4), 579-619. doi: 10.2307/1170654.

[26] Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*(3), 519-530. doi: 10.2307/2334770.

[27] Maxwell, S. E. (1980). Pairwise multiple comparisons in repeated measures designs. *Journal of Educational and Behavioral Statistics, 5*(3), 269-287. doi: 10.3102/10769986005003269.

[28] Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, N.J.: Lawrence Erlbaum Associates.

[29] Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166.

[30] Oberfeld, D., & Franke, T. (2013). Evaluating the robustness of repeated measures analyses: The case of small sample sizes and non-normal data. *Behavior Research Methods, 45*(3), 792-812. doi: http:/dx.doi.org/10.3758/s13428-012-0281-2.

[31] Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434-447. doi: 10.1037/1082-989x.8.4.434.

[32] Rasmussen, J. L. (1987). Parametric and Bootstrap Approaches to Repeated Measures Designs. *Behavior Research Methods Instruments & Computers, 19*(4), 357-360.

[33] Rouanet, H., & Lépine, D. (1970). Comparison between treatments in a repeated-measurement design: ANOVA and multivariate methods. *British Journal of Mathematical and Statistical Psychology, 23*(2), 147-163.

[34] Royston, J. P. (1983). Some techniques for assessing multivariate normality based on the Shapiro-Wilk-W. *Applied Statistics-Journal of the Royal Statistical Society Series C, 32*(2), 121-133. doi: 10.2307/2347291.

[35] Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology-European Journal of Research Methods for the Behavioral and Social Sciences, 6*(4), 147-151. doi: 10.1027/1614-2241/a000016.

[36] St. Laurent, R., & Turk, P. (2013). The effects of misconceptions on the properties of Friedman's test. *Communications in Statistics-Simulation and Computation, 42*(7), 1596-1615. doi: 10.1080/03610918.2012.671874.

[37] Tukey, J. W. (1977). Exploratory data analysis. Reading, Mass.: Addison-Wesley Pub. Co.

[38] Seco, G. V., Izquierdo, M. C., García, M. P. F., & Díez, F. J. H. (2006). A comparison of the bootstrap-F, improved general approximation, and Brown-Forsythe multivariate approaches in a mixed repeated measures design. *Educational and Psychological Measurement, 66*(1), 35-62.

[39] Wilcox, R. R., Keselman, H. J., Muska, J., & Cribbie, R. (2000). Repeated measures ANOVA: Some new results on comparing trimmed means and means. *British Journal of Mathematical & Statistical Psychology, 53*, 69-82.

附件1的
附录 5
(参考性)

最佳锚点特性要求

使任何成功的锚点被设计得最适宜捕获的关键描述符给出如下。

最佳锚点特性应该：

1) 当与采用ITU-R BS.1534建议书中锚点规范采集到的数据相比较时，产生在测试系统相对排序中不出现实质性变化的数据；
2) 当与采用ITU-R BS.1534建议书中锚点规范为被测系统收集的数据相比较时，与对测试系统使用一个更宽范围评分量表的收听者评分相关联；
3) 被收听者感觉为对被测系统比对ITU-R BS.1534建议书所描述的锚点更熟悉。这可能反过来导致更长的锚点评估时间。
4) 使得能够对中等范围测试系统进行敏感的比较；
5) 在低范围和中等范围锚点之间产生大约20-30点的评分差；
6) 在具有有限内容相关性的锚点中产生质量损伤。